Rademacher Chaos Complexities for Learning the Kernel Problem

Yiming Ying

mathying@gmail.com

College of Engineering, Mathematics and Physical Sciences

University of Exeter, Harrison Building

Exeter, EX4 4QF, UK

Colin Campbell

C.Campbell@bristol.ac.uk

Department of Engineering Mathematics

University of Bristol, Queen's Building

Bristol, BS8 1TR, UK

Abstract

In this paper we develop a novel generalization bound for learning the kernel problem. First, we show that the generalization analysis of the kernel learning problem reduces to investigation of the suprema of the Rademacher chaos process of order two over candidate kernels, which we refer to as *Rademacher chaos complexity*. Next, we show how to estimate the empirical Rademacher chaos complexity by well-established metric entropy integrals and pseudo-dimension of the set of candidate kernels. Our new methodology mainly depends on the principal theory of U-processes and entropy integrals. Finally, we establish satisfactory excess generalization bounds and misclassification error rates for learning Gaussian kernels and general radial basis kernels.

Keywords: Learning the kernel, generalization bound, Rademacher chaos, Rademacher averages, entropy integrals

1 Introduction

Kernel methods such as Support Vector Machines (SVM) have been extensively applied to supervised learning tasks such as classification and regression, see e.g. Schölkopf and Smola (2002); Shawe-Taylor and Cristianini (2004); Cucker and Zhou (2007); Steinwart and Christmann (2008). The performance of a kernel machine largely depends on the data representation via the choice of kernel function. Hence, one central issue in kernel methods is the problem of kernel selection.

Kernel learning can range from the width parameter selection of Gaussian kernels to obtaining an optimal linear combination from a set of finite candidate kernels. The latter is often referred to as *multiple kernel learning (MKL)* in Machine Learning and *non-parametric Group Lasso* (Bach, 2008) in Statistics. Lanckriet et al. (2004) pioneered work on MKL and proposed to automatically learn a linear combination of candidate kernels for the case of SVMs using a semi-definite programming (SDP) approach. Similar problems studied recently include hyperkernels (Ong and Smola, 2005), Bayesian probabilistic kernel learning models (Girolami and Rogers, 2005), kernel discriminant analysis (Ye et al., 2008) and information-thereotic data integration (Ying et al., 2009) etc. Such MKL formulations have been successfully demonstrated in combining multiple heterogeneous data sources to enhance biological inference (Lanckriet et al., 2004; Damoulas and Girolami, 2008; Ying et al., 2009).

The above mentioned MKL models usually learn an optimal combination from a finite set of candidate kernels. A general regularization framework including kernel hyper-parameter learning and MKL was formulated in Micchelli and Pontil (2005); Wu et al. (2006) with a potentially *infinite* number of candidate kernels which is generally referred to as the *learning the kernel problem*. Specifically, let $\mathbb{N}_n = \{1, 2, \cdots, n\}$ for any $n \in \mathbb{N}$ and we are interested in the classification problem on the input space $X \subseteq \mathbb{R}^d$ and output space $Y = \{\pm 1\}$. The relation between input X and output Y is specified by a set of training samples $\mathbf{z} = \{z_i = (x_i, y_i) : x_i \in X, \ y_i \in Y, \ i \in \mathbb{N}_n\}$ which are identically and independently distributed (i.i.d.) according to an unknown distribution ρ on $Z = X \times Y$. Let $\mathcal{K}$ be a prescribed (possible infinite) set of candidate (base) kernels and denote the candidate reproducing kernel Hilbert space (RKHS) with kernel K by $\mathcal{H}_K$ with norm $\|\cdot\|_K$. In addition, we always assume that the quantity $\kappa :=$

$\sup_{K \in \mathcal{K}, x \in X} \sqrt{K(x,x)}$ is finite. Then the general kernel learning scheme (Micchelli and Pontil, 2005; Wu et al., 2006) can be cast as a two-layer minimization problem:

$$f_{\mathbf{z}}^{\phi} = \arg \min_{K \in \mathcal{K}} \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{n} \sum_{i \in \mathbb{N}_n} \phi\big(y_i f(x_i)\big) + \lambda \|f\|_K^2 \right\}. \tag{1}$$

Here, $\phi : \mathbb{R} \to [0, \infty)$ is a loss function for classification and λ is a positive regularization parameter. We use the superscript ϕ of $f_{\mathbf{z}}^{\phi}$ to emphasize the solution $f_{\mathbf{z}}^{\phi}$ is produced by scheme (1) with loss function ϕ. When the loss function ϕ is the hinge loss and $\mathcal{K}$ is the linear combination of the set of finite base kernels $\{K_\ell : \ell \in \mathbb{N}_m\}$, i.e. $\mathcal{K} := \{\sum_{\ell \in \mathbb{N}_m} \lambda_\ell \mathcal{K}_\ell : \sum_{\ell \in \mathbb{N}_m} \lambda_\ell = 1, \lambda_\ell \geq 0, \ \forall \ell \in \mathbb{N}_m\}$, then the above kernel learning framework (1) is reduced to the standard margin-based MKL formulation (Lanckriet et al., 2004). If $\mathcal{K} = \{e^{-\sigma \|x-t\|^2} : \sigma > 0\}$ then it is reduced to the formulation for learning the Gaussian kernel hyper-parameter (Chapelle et al., 2002).

Statistical generalization analysis of learning the kernel problem (1) was pursued by Bousquet and Herrmann (2003); Lanckriet et al. (2004); Ying and Zhou (2007); Micchelli et al. (2005); Srebro and Ben-David (2006). In this paper we leverage Rademacher complexity bounds for empirical risk minimization (ERM) and for SVM with a single kernel (Bartlett and Mendelson, 2002; Bartlett et al., 2006; Koltchinskii and Panchenko, 2002) and develop a novel generalization bound for kernel learning problem (1). In particular, we show that generalization analysis of the kernel learning algorithms reduces to investigation of the suprema of the Rademacher chaos process of order two over candidate kernels, which we refer to as *Rademacher chaos complexity*. Next, we show how to estimate the empirical Rademacher chaos complexity by well-established metric entropy integrals and pseudo-dimension of the set of candidate kernels. Our new methodology mainly depends on the principal theory of U-processes (De La Peña and Giné, 1999). A preliminary version of this paper has appeared in the COLT conference Ying and Colin (2009).

This paper is organized as follows. In Section 2 we illustrate our main theorems. The main proofs for theorems are given in Section 3 and Section 4. Explicit error rates with examples for learning Gaussian kernels and radial basis kernels are given in Section 5. In Section 6 we discusses related work and compares our results with those in the literature. The last section concludes the paper.

2 Main Results

In this section we illustrate our main contributions.

2.1 Main Theorems

The true error or *generalization error* is defined as

$$\mathcal{E}^\phi(f) = \iint_{X \times Y} \phi(yf(x))d\rho(x, y),$$

and the target function f_ρ^ϕ is defined by $f_\rho^\phi = \arg\min_f \mathcal{E}^\phi(f)$. Let the empirical error $\mathcal{E}_{\mathbf{z}}$ be defined, for any f, by

$$\mathcal{E}_{\mathbf{z}}^\phi(f) = \frac{1}{n}\sum_{j \in \mathbb{N}_n} \phi(y_j f(x_j)).$$

For brevity, throughout this paper we restrict our interest to a large class of loss functions for classification (Wu et al., 2006), see also a general definition of classification loss functions in Bartlett et al. (2006).

Definition 1 *A function $\phi : \mathbb{R} \to [0, \infty)$ is called a normalized classifying loss if it is convex, $\phi'(0) < 0$, $\inf_{t \in \mathbb{R}} \phi(t) = 0$, and $\phi(0) = 1$.*

Our target is to bound the true error by the empirical error. To this end, let the union of the unit ball of candidate RKHSs be denoted by

$$\mathcal{B}_\mathcal{K} := \Big\{ f : \ f \in \mathcal{H}_K \text{ and } \|f\|_K \leq 1, \ K \in \mathcal{K} \Big\}.$$

By the definition of $f_{\mathbf{z}}^\phi$, we get, for some RKHS $\mathcal{H}_K$, that $\frac{1}{n}\sum_{i=1}^n \phi\big(y_i f_{\mathbf{z}}^\phi(x_i)\big) + \lambda\|f_{\mathbf{z}}^\phi\|_K^2 \leq \frac{1}{n}\sum_{i=1}^n \phi(0) + \lambda\|0\|_K^2 = 1$. Hence, $\|f_{\mathbf{z}}^\phi\|_K \leq \sqrt{1/\lambda}$. This implies, for any samples $\mathbf{z}$, that

$$f_{\mathbf{z}}^\phi \in \mathcal{B}_\lambda := \frac{1}{\sqrt{\lambda}} \mathcal{B}_\mathcal{K} := \Big\{ \frac{f}{\sqrt{\lambda}} : f \in \mathcal{B}_\mathcal{K} \Big\}. \tag{2}$$

Hence, $\|f_{\mathbf{z}}^\phi\|_\infty < \kappa\sqrt{1/\lambda}$. Finally, for a Lipschitz continuous function $\psi : \mathbb{R} \to [0, \infty)$ we need the constant defined by

$$M_\lambda^\psi = \sup \Big\{ |\psi(t)| : \ \forall |t| \leq \kappa\sqrt{1/\lambda} \Big\}, \tag{3}$$

4

and denote its local Lipschitz constant by

$$C_\lambda^\psi = \sup\left\{\frac{|\psi(x) - \psi(x')|}{|x - x'|} \; : \; \forall |x|, |x'| \leq \kappa\sqrt{\frac{1}{\lambda}}\right\}. \tag{4}$$

If $\psi = \phi$ is convex, then ϕ's left derivative ϕ'_- and right one ϕ'_+ are well defined and C_λ^ϕ is identical to $C_\lambda^\phi = \sup\{\max(|\phi'_-(t)|, |\phi'_+(t)|) \; : \; \forall |t| \leq \kappa\sqrt{1/\lambda}\}$.

Our generalization analysis depends on the suprema of the *homogeneous Rademacher chaos of order two* over a class of functions defined as follows, see Chapter 3.2 of De La Peña and Giné (1999) for a general definition of Rademacher chaos of order m for any $m \in \mathbb{N}$.

Definition 2 *Let F be a class of functions on $X \times X$ and $\{\epsilon_i : i \in \mathbb{N}_n\}$ are independent Rademacher random variables. Also, $\mathbf{x} = \{x_i : i \in \mathbb{N}_n\}$ are independent random variables distributed according to a distribution μ on X. The homogeneous Rademacher chaos process of order two, with respect to the Rademacher variable ε, is a random variable system defined by $\{\hat{U}_f(\varepsilon) = \frac{1}{n}\sum_{i,j\in\mathbb{N}_n, i<j} \varepsilon_i\varepsilon_j f(x_i, x_j) : f \in F\}$. We refer to the expectation of its suprema*

$$\hat{\mathcal{U}}_n(F) = \mathbb{E}_\varepsilon[\sup_{f\in F}|\hat{U}_f(\varepsilon)|]$$

as the empirical Rademacher chaos complexity over F.

It is worth mentioning that the Rademacher process $\{\frac{1}{\sqrt{n}}\sum_{i\in\mathbb{N}_n} \varepsilon_i f(x_i) : f \in F\}$ for Rademacher averages can be regarded as a *homogeneous Rademacher chaos process of order one*. The nice application of U-processes to the generalization analysis of ranking and scoring problem is recently developed in Clémencon et al. (2008).

Our first main result shows that the excess generalization error of MKL algorithms can be bounded by the empirical Rademacher chaos complexity over the set of candidate kernels.

Theorem 1 *Let ϕ be a normalized classifying loss. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, there holds*

$$\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}_{\mathbf{z}}^\phi(f_{\mathbf{z}}^\phi) \leq 2C_\lambda^\phi\left(\frac{2\hat{\mathcal{U}}_n(\mathcal{K})}{\lambda n}\right)^{\frac{1}{2}} + 2\kappa C_\lambda^\phi\left(\frac{1}{n\lambda}\right)^{\frac{1}{2}} + 3M_\lambda^\phi\left(\frac{\ln(\frac{2}{\delta})}{n}\right)^{\frac{1}{2}}. \tag{5}$$

Theorem 1 will be proved in Section 3.

Now we apply the well-established theory of U processes to estimate Rademacher chaos complexity by entropy integrals. To this end, let $\mathcal{G}$ be a set of functions on $X \times X$ and $\mathbf{x} = \{x_i \in X : i \in \mathbb{N}_n\}$, define the l^2 empirical metric of two functions $f, g \in \mathcal{G}$ by

$$d_\mathbf{x}(f, g) = \left(\frac{1}{n^2} \sum_{i,j \in \mathbb{N}_n, i < j} |f(x_i, x_j) - g(x_i, x_j)|^2 \right)^{\frac{1}{2}}.$$

The empirical covering number $\mathcal{N}(\mathcal{G}, d_\mathbf{x}, \eta)$ is the smallest number of balls with radius η required to cover $\mathcal{G}$. The empirical Rademacher chaos complexity $\hat{\mathcal{U}}_n(\mathcal{K})$ can be bounded by the metric entropy integral as follows.

Theorem 2 *For any* $\mathbf{x} = \{x_i : i \in \mathbb{N}_n\}$, *there holds*

$$\hat{\mathcal{U}}(\mathcal{K}) \leq \kappa^2 + 12e \int_0^{\kappa^2} \log\left[1 + \mathcal{N}(\mathcal{K}, d_\mathbf{x}, \delta) \right] d\delta$$

Theorem 2 will be proved in Section 4. Theorem 2 suggests that if $\log \mathcal{N}(\mathcal{K}, d_\mathbf{x}, \varepsilon) = \mathcal{O}(\varepsilon^{-p})$ with some $0 \leq p < 1$ then the Rademacher chaos complexity $\hat{\mathcal{U}}_n(\mathcal{K})$ is uniformly bounded. To estimate the covering number, a simple case would bound it by the number of candidate kernels. For example, if

$$\mathcal{K}_{\text{finite}} = \{K_\ell : \ell \in \mathbb{N}_m\} \tag{6}$$

then $\mathcal{N}(\mathcal{K}, d_\mathbf{x}, \varepsilon) \leq m$ and hence

$$\hat{\mathcal{U}}_n(\mathcal{K}_{\text{finite}}) \leq \kappa^2 + 12e\kappa^2 \log(m+1) \leq 24e\kappa^2 \log(m+1), \ \forall m \geq 2. \tag{7}$$

If the candidate kernel set has an infinite number of kernels, the covering number can further be estimated by capacity numbers such as the pseudo-dimension. For this purpose, we recall the definition of kernel pseudo-dimension of a class of kernel functions on the product space $X \times X$, see Anthony and Bartlett (1999).

Definition 3 *Let* $\mathcal{K}$ *be a set of reproducing kernel functions mapping from* $X \times X$ *to* $\mathbb{R}$. *We say that* $S_m = \{(x_i, t_i) \in X \times X : i \in \mathbb{N}_m\}$ *is pseudo-shattering by* $\mathcal{K}$ *if there are real numbers* $\{r_i \in \mathbb{R} : i \in \mathbb{N}_m\}$ *such that for any* $b \in \{-1, 1\}^m$ *there is a function* $K \in \mathcal{K}$ *with property* $sgn(K(x_i, t_i) - r_i) = b_i$ *for any* $i \in \mathbb{N}_m$. *Then, we define a pseudo-dimension* $d_\mathcal{K}$ *of* $\mathcal{K}$ *to be the maximum cardinality of* S_m *that is pseudo-shattered by* $\mathcal{K}$.

We are now ready to estimate the Rademacher chaos complexity using pseudo-dimensions.

Theorem 3 *If the pseudo-dimension $d_{\mathcal{K}}$ of the set of basis kernels is finite, then we have that*

$$\mathcal{N}(\mathcal{K}, d_{\mathbf{x}}, \varepsilon) \leq 2\Big(\frac{4e\kappa^4}{\varepsilon^2}\Big)^{d_{\mathcal{K}}}. \tag{8}$$

Moreover, for any $\mathbf{x} = \{x_i : i \in \mathbb{N}_n\}$, there holds

$$\hat{\mathcal{U}}_n(\mathcal{K}) \leq (96e + 1)\kappa^2 d_{\mathcal{K}}. \tag{9}$$

Theorem 3 will be proved in Section 4. For Gaussian-type kernels, we can explicitly estimate the pseudo-dimension, and hence bound the empirical Rademacher chaos complexities. To see this, consider the set of scalar candidate kernels given by

$$\mathcal{K}_{\text{gau}} = \big\{e^{-\sigma\|x-t\|^2} : \ \sigma \in [0, \infty)\big\}. \tag{10}$$

The second class of candidate kernels is more general as considered in Micchelli et al. (2005): the whole class of *radial basis kernels*. Let $\mathcal{M}(\mathbb{R}^+)$ be the class of probabilities on $\mathbb{R}^+$. We consider the candidate kernel defined by

$$\mathcal{K}_{\text{rbf}} = \Big\{ \int_0^\infty e^{-\sigma\|x-t\|^2} dp(\sigma) : p \in \mathcal{M}(\mathbb{R}^+) \Big\} \tag{11}$$

Overall, for the above specific sets of basis kernels, we can have the following result.

Corollary 1 *For the Rademacher chaos complexity of $\mathcal{K}$, we respectively have the following estimation:*

1. *If $\mathcal{K}$ has a finite number of kernels given by (6) then*

$$\hat{\mathcal{U}}_n(\mathcal{K}_{\text{finite}}) \leq 24e\kappa^2 \log(m+1).$$

2. *If K is the set of gaussian-type kernels given by equation (10) and (11) then*

$$\hat{\mathcal{U}}_n(\mathcal{K}_{\text{rbf}}) \leq \hat{\mathcal{U}}_n(\mathcal{K}_{\text{gau}}) \leq (1 + 96e)\kappa^2.$$

Corollary 1 will be proved in Section 4. Combining Theorem 1 with Corollary 1, the generalization bound can be summarized as follows: with probability at least $1 - \delta$ there holds

$$\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}_{\mathbf{z}}^\phi(f_{\mathbf{z}}^\phi) \leq 4\left(C_\lambda^\phi \Big(\frac{(192e+2)\kappa^2 d_{\mathcal{K}}}{n\lambda}\Big)^{\frac{1}{2}} + M_\lambda^\phi \big(\frac{\ln\frac{2}{\delta}}{n}\big)^{\frac{1}{2}}\right). \tag{12}$$

Moveover, if $\mathcal{K} = \mathcal{K}_{\text{finite}}$ is given by equation (6) then the above bound is reduced to

$$\mathcal{E}^\phi(f_\mathbf{z}^\phi) - \mathcal{E}_\mathbf{z}^\phi(f_\mathbf{z}^\phi) \leq 4\left(C_\lambda^\phi\left(\frac{48e\kappa^2\ln(m+1)}{n\lambda}\right)^{\frac{1}{2}} + M_\lambda^\phi(\frac{\ln\frac{2}{\delta}}{n})^{\frac{1}{2}}\right). \tag{13}$$

where m is the number of candidate kernels in $\mathcal{K}$.

We conclude this subsection with an important remark on the bounds for learning a convex hull of candidate kernels. All the above estimations and bounds for the Rademacher chaos complexity hold true for the convex hull of $\mathcal{K}$ defined by

$$\text{conv}(\mathcal{K}) := \left\{\sum_{j\in\mathbb{N}_p}\lambda_\ell K_\ell : K_\ell \in \mathcal{K}, \lambda_\ell \geq 0, \sum_{\ell\in\mathbb{N}_p}\lambda_\ell = 1, p \in \mathbb{N}\right\},$$

since it is easy to check, by the definition of the Rademacher chaos complexity, that

$$\hat{\mathcal{U}}_n\big(\text{conv}(\mathcal{K})\big) \leq \hat{\mathcal{U}}_n(\mathcal{K}).$$

2.2 Error Rates in Classification

In this subsection we derive misclassification error rates for *multi-kernel regularized classifier* $\text{sign}(f_\mathbf{z}^\phi)$ with the regularization scheme (1). The quality of a classifier $\mathcal{C} : X \to Y$ is measured by the *misclassification error* which is defined by

$$\mathcal{R}(\mathcal{C}) := \iint_{X\times Y} P(y \neq \mathcal{C}(x)|x)d\rho(x,y). \tag{14}$$

The target is to understand how $\text{sign}(f_\mathbf{z}^\phi)$ approximates the Bayes rule f_c (Devroye et al., 1997) defined by $f_c = \arg\inf \mathcal{R}(\mathcal{C})$. More specifically, we aim to estimate the *excess misclassification error*

$$\mathcal{R}(\text{sign}(f_\mathbf{z}^\phi)) - \mathcal{R}(f_c).$$

As shown in Zhang (2004); Bartlett et al. (2006), the excess misclassification error can usually be bounded by the *excess generalization error*: $\mathcal{E}^\phi(f_\mathbf{z}^\phi)) - \mathcal{E}(f_\rho^\phi)$. To transfer generalization bounds in Subsection 2.1 to the misclassificaiton error bounds, we introduce the error decomposition of problem (1).

Let the empirical error $\mathcal{E}_\mathbf{z}$ be defined, for any f, by $\mathcal{E}_\mathbf{z}^\phi(f) = \frac{1}{n}\sum_{j\in\mathbb{N}_n}\phi(y_jf(x_j))$. We also introduce the *regularization error* defined by

$$\mathcal{D}(\lambda) = \inf_{K\in\mathcal{K}}\inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}^\phi(f) - \mathcal{E}^\phi(f_\rho^\phi) + \lambda\|f\|_K^2\right\}$$

and call the minimizer f_λ^ϕ of the regularization error the *regularization function*. In addition, we define the *sample error* $\mathcal{S}_{\mathbf{z},\lambda}$ by

$$\mathcal{S}_{\mathbf{z},\lambda} = \left\{ \mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}_{\mathbf{z}}^\phi(f_{\mathbf{z}}^\phi) \right\} + \left\{ \mathcal{E}_{\mathbf{z}}^\phi(f_\lambda^\phi) - \mathcal{E}^\phi(f_\lambda^\phi) \right\}.$$

From the standard *error decomposition* (Zhang, 2004; Bartlett et al., 2006; Steinwart and Scovel, 2007; Ying and Zhou, 2007), we have that

$$\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}^\phi(f_\rho^\phi) \leq \mathcal{D}(\lambda) + \mathcal{S}_{\mathbf{z},\lambda}. \tag{15}$$

Throughout this paper, for simplicity we always assume the existence of the empirical solution $f_{\mathbf{z}}^\phi$ and the regularization function f_λ^ϕ, see discussions in Appendix B of Ying and Zhou (2007).

We are now ready to state misclassification error rates. Hereafter, the expression $a_n = \mathcal{O}(b_n)$ means that there exists an absolute constant c such that $a_n \leq cb_n$ for all $n \in \mathbb{N}$. We usually assume conditions on the distribution ρ or some regularity condition on the target function f_ρ^ϕ under which the regularization error $\mathcal{D}(\lambda)$ decays polynomially. For instance, we can employ the following condition.

Definition 4 *We say that ρ is separable by $\{\mathcal{H}_K : K \in \mathcal{K}\}$ if there is some $f_{sp} \in \mathcal{H}_{\bar{K}}$ with some $\bar{K} \in \mathcal{K}$ such that $y f_{sp}(x) > 0$ almost surely. It has separation exponent $\theta \in (0, \infty]$ if we can choose f_{sp} and positive constants Δ, c_θ such that $\|f_{sp}\|_{\bar{K}} = 1$ and*

$$\rho_X \left\{ x \in X : |f_{sp}(x)| < \Delta t \right\} \leq c_\theta t^\theta, \qquad \forall t > 0. \tag{16}$$

Observe that condition (16) with $\theta = \infty$ is equivalent to

$$\rho_X \left\{ x \in X : |f_{\mathrm{sp}}(x)| < \gamma t \right\} = 0, \qquad \forall\, 0 < t < 1.$$

That is, $|f_{\mathrm{sp}}(x)| \geq \gamma$ almost everywhere. Thus, separable distributions with separation exponent $\theta = \infty$ correspond to strictly separable distributions. Other assumptions on the distribution ρ such as the geometric noise condition introduced in Steinwart and Scovel (2005) are possible to achieve polynomial decays of the regularization error.

Example 1 *Let $\phi(t) = (1 - t)_+$ be the hinge loss and consider the MKL algorithm (1) with $\mathcal{K}$ given by either $\mathcal{K}_{sc}$ or $\mathcal{K}_{rbf}$. Suppose that the separation condition holds*

true with exponent $\theta > 0$. Then, by choosing $\lambda = n^{-\frac{2+\theta}{(2+3\theta)}}$, for any $\delta \in (0,1)$, with probability at least $1 - \delta$ there holds

$$\mathcal{R}(sgn(f_{\mathbf{z}}^{\phi})) - \mathcal{R}(f_c) \le \mathcal{O}\Big(\big[\ln(1/\delta)\big]^{\frac{1}{2}}\Big(\frac{1}{n}\Big)^{\frac{\theta}{3\theta+2}}\Big).$$

The proof of this example is postponed to Section 5. Other examples such as least square loss regression can be found in Section 5. In this case we need to consider the function approximation (De Vito et al., 2006; Smale and Zhou, 2004; Ye and Zhou, 2008) on a domain or low dimensional manifold of $\mathbb{R}^d$.

In analogy to the data-dependent risk bounds of Rademacher averages (Bartlett et al., 2006), we can get margin bounds for learning the kernel problems using Rademacher chaos complexities.

Corollary 2 *Let $\phi(t) = (1-t)_+$ be the hinge loss and $\gamma > 0$, $0 < \delta < 1$ and define the margin cost function by*

$$\psi(t) = \begin{cases} 1, & t \le 0 \\ 1 - \frac{t}{\gamma}, & 0 < t \le \gamma \\ 0, & t > \gamma \end{cases} \tag{17}$$

Then, with probability at least $1 - \delta$, there holds

$$\mathcal{R}(sgn(f_{\mathbf{z}}^{\phi})) \le \mathcal{E}_{\mathbf{z}}^{\psi}(f_{\mathbf{z}}^{\phi}) + 2\Big(\frac{2\hat{\mathcal{U}}_n(\mathcal{K})}{n\lambda\gamma^2}\Big)^{\frac{1}{2}} + 2\kappa\Big(\frac{1}{n\lambda\gamma^2}\Big)^{\frac{1}{2}} + 3\Big(\frac{\ln(\frac{2}{\delta})}{n}\Big)^{\frac{1}{2}}.$$

Corollary 2 will be proved in Section 5. When $\mathcal{K}$ only has a single kernel K, we have

$$\begin{aligned} \hat{\mathcal{U}}_n(K) &\le \mathbb{E}_{\varepsilon}\Big|\frac{1}{n}\sum_{i,j\in\mathbb{N}_n}\varepsilon_i\varepsilon_j K(x_i,x_j)\Big| + \Big|\frac{1}{n}\sum_{i\in\mathbb{N}_n}K(x_i,x_i)\Big| \\ &= \mathbb{E}_{\varepsilon}\frac{1}{n}\sum_{i,j\in\mathbb{N}_n}\varepsilon_i\varepsilon_j K(x_i,x_j) + \frac{1}{n}\sum_{i\in\mathbb{N}_n}K(x_i,x_i) \end{aligned}$$

where the last equality follows from the positive semi-definiteness of kernel K. Hence, the Rademacher chaos complexity can be estimated by

$$\hat{\mathcal{U}}_n(K) \le \frac{2}{n}\sum_{i\in\mathbb{N}_n}K(x_i,x_i) := \frac{2}{n}\text{trace}(\mathbf{K}),$$

where $\mathbf{K} = (K(x_i,x_j))_{i,j\in\mathbb{N}_n}$. Consequently, Corollary 2 implies that

$$\mathcal{R}(sgn(f_{\mathbf{z}}^{\phi})) \le \mathcal{E}_{\mathbf{z}}^{\psi}(f_{\mathbf{z}}^{\phi}) + \frac{4}{\gamma}\frac{\sqrt{\text{trace}(\mathbf{K})}}{n\sqrt{\lambda}} + 2\kappa\Big(\frac{1}{n\lambda\gamma^2}\Big)^{\frac{1}{2}} + 3\Big(\frac{\ln(\frac{2}{\delta})}{n}\Big)^{\frac{1}{2}}.$$

This coincides with the bound in Bartlett and Mendelson (2002) for the single kernel case with solutions $f_{\mathbf{z}}^{\phi}$ in the function space $\big\{f = \sum_{i\in\mathbb{N}_n} \alpha_i K(x_i,\cdot) : \|f\|_K \leq \frac{1}{\sqrt{\lambda}}\big\}$.

We now present an example of margin bounds which can be directly obtained by combining Corollary 1 with Corollary 2.

Example 2 *Let $\phi(t) = (1-t)_+$ be the hinge loss. Then, for any margin $\gamma > 0$, we have the following estimation for gaussian-type kernel set and the set of finite kernels:*

1. If $\mathcal{K} = \mathcal{K}_{gau}$ or $\mathcal{K} = \mathcal{K}_{rbf}$ then, with probability $1-\delta$, there holds

$$\mathcal{R}(sgn(f_{\mathbf{z}}^{\phi})) \leq \mathcal{R}_{\mathbf{z}}^{\gamma}(f_{\mathbf{z}}^{\phi}) + 2\Big(\frac{(192e+2)\kappa^2}{n\lambda\gamma^2}\Big)^{\frac{1}{2}} + 2\kappa\Big(\frac{1}{n\lambda\gamma^2}\Big)^{\frac{1}{2}} + 3\Big(\frac{\ln\frac{2}{\delta}}{n}\Big)^{\frac{1}{2}}.$$

2. If $\mathcal{K}$ is the convex hull of m candidate kernels, then, with probability $1-\delta$,

$$\mathcal{R}(sgn(f_{\mathbf{z}}^{\phi})) \leq \mathcal{R}_{\mathbf{z}}^{\gamma}(f_{\mathbf{z}}^{\phi}) + 2\Big(\frac{48e\kappa^2 \log(m+1)}{n\lambda\gamma^2}\Big)^{\frac{1}{2}} + 2\kappa\Big(\frac{1}{n\lambda\gamma^2}\Big)^{\frac{1}{2}} + 3\Big(\frac{\ln\frac{2}{\delta}}{n}\Big)^{\frac{1}{2}}.$$

3 Generalization Bounds by Rademacher Chaos

In this section we prove Theorem 1 which states that the generalization bound of MKL algorithm (1) can be bounded by well-established Rademacher chaos of order two. To this end, recall the definition of the ordinary *Rademacher averages*, see e.g. Bartlett and Mendelson (2002); Bartlett et al. (2005); Koltchinskii (2001); Koltchinskii and Panchenko (2002).

Definition 5 *Let μ be a probability measure on Ω and F be a class of uniformly bounded and measurable functions on Ω. For any $n \in \mathbb{N}$, define the random variable by*

$$\hat{R}_n(F) := \frac{1}{\sqrt{n}} \sup_{f\in F} \Big| \sum_{i\in\mathbb{N}_n} \epsilon_i f(z_i)\Big|$$

where $\{z_i : i \in \mathbb{N}_n\}$ are independent random variables distributed according to μ and $\{\epsilon_i : i = 1,\ldots,n\}$ are independent Rademacher random variables, that is, $P(\epsilon_i = +1) = P(\epsilon_i = -1) = 1/2$. Also, we often call $R_n(F) := \mathbb{E}[\hat{R}_n(F)] = \mathbb{E}_\mu \mathbb{E}_\varepsilon[R_n(F)]$ the Rademacher averages (complexity)[1] over the class F.

---

[1]The empirical Rademacher average is usually defined by $\hat{R}_n(F) := \frac{1}{n}\sup_{f\in F}\big|\sum_{i\in\mathbb{N}_n}\epsilon_i f(z_i)\big|$. For technical simplicity, we use its scaling version here.

Hence, $\hat{R}_n(F)$ is the suprema of the Rademacher process $\{\frac{1}{\sqrt{n}} \sum_{i \in \mathbb{N}_n} \epsilon_i f(z_i) : f \in F\}$ indexed by F which can also be regarded as the homogenous Rademacher chaos process of order one. Some useful properties of Rademacher averages can be found in Bartlett and Mendelson (2002). Now we assemble the necessary materials to obtain the main technical result.

Theorem 4 *Let ψ be a Lipschitz continuous cost function satisfying $\inf_t \psi(t) = 0$ and $\psi(0) = 1$. Let $\mathcal{B}_\lambda$ be defined by equation (2) and M_λ^ψ, $\mathcal{C}_\lambda^\psi$ be respectively defined by (3) and (4). Then, with probability at least $1 - \delta$, there holds*

$$\sup_{f \in \mathcal{B}_\lambda} \left[\mathcal{E}^\psi(f) - \mathcal{E}_{\mathbf{z}}^\psi(f) \right] \leq 2C_\lambda^\psi \left(\frac{2\hat{\mathcal{U}}_n(\mathcal{K})}{\lambda n} \right)^{\frac{1}{2}} + 2\kappa C_\lambda^\psi \left(\frac{1}{n\lambda} \right)^{\frac{1}{2}} + 3M_\lambda^\psi \left(\frac{\ln(\frac{2}{\delta})}{n} \right)^{\frac{1}{2}}.$$

Similarly, with probability at least $1 - \delta$, we have

$$\sup_{f \in \mathcal{B}_\lambda} \left[\mathcal{E}_{\mathbf{z}}^\psi(f) - \mathcal{E}^\psi(f) \right] \leq 2C_\lambda^\psi \left(\frac{2\hat{\mathcal{U}}_n(\mathcal{K})}{\lambda n} \right)^{\frac{1}{2}} + 2\kappa C_\lambda^\psi \left(\frac{1}{n\lambda} \right)^{\frac{1}{2}} + 3M_\lambda^\psi \left(\frac{\ln(\frac{2}{\delta})}{n} \right)^{\frac{1}{2}}.$$

Proof By McDiarmid's bounded difference inequality (see e.g. Devroye et al. (1997)), with probability $1 - \frac{\delta}{2}$ we have that

$$\sup_{f \in \mathcal{B}_\lambda} \left[\mathcal{E}^\psi(f) - \mathcal{E}_{\mathbf{z}}^\psi(f) \right] \leq \mathbb{E} \sup_{f \in \mathcal{B}_\lambda} \left[\mathcal{E}^\psi(f) - \mathcal{E}_{\mathbf{z}}^\psi(f) \right] + M_\lambda^\psi \left(\frac{\ln \frac{2}{\delta}}{2n} \right)^{\frac{1}{2}}. \tag{18}$$

The first term on the righthand side of the above inequality can be estimated by the standard symmetrization arguments. Indeed, with probability at least $1 - \frac{\delta}{2}$, there holds

$$\begin{aligned}
\mathbb{E}\left[\sup_{f \in \mathcal{B}_\lambda} \left(\mathcal{E}^\psi(f) - \mathcal{E}_{\mathbf{z}}^\psi(f) \right) \right] &\leq 2\mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{B}_\lambda} \frac{1}{n} \sum_{i \in \mathbb{N}_n} \epsilon_i \psi(y_i f(x_i)) \right] \\
&\leq 2\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{B}_\lambda} \frac{1}{n} \sum_{i \in \mathbb{N}_n} \epsilon_i \psi(y_i f(x_i)) \right] + 2M_\lambda^\psi \left(\frac{\ln \frac{2}{\delta}}{2n} \right)^{\frac{1}{2}},
\end{aligned} \tag{19}$$

where the last inequality used again the McDiarmid's bounded difference inequality. Note that $\|f\|_\infty \leq \kappa \sqrt{1/\lambda}$ for any $f \in \mathcal{B}_\lambda$. Then, from the definition of C_λ^ψ given by equation (4), function ψ has a Lipschitz constant C_λ^ψ. Applying the contraction property of Rademacher averages (Theorem 7 of Meir and Zhang (2003)) implies that, with probability $1 - \frac{\delta}{2}$,

$$\begin{aligned}
\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{B}_\lambda} \sum_{i \in \mathbb{N}_n} \epsilon_i \psi(y_i f(x_i)) \right] &\leq \mathbb{E}_\varepsilon \sup_{f \in \mathcal{B}_\lambda} \sum_{i \in \mathbb{N}_n} \epsilon_i \psi(y_i f(x_i)) \\
&\leq C_\lambda^\psi \, \mathbb{E}_\varepsilon \sup_{f \in \mathcal{B}_\lambda} \sum_{i \in \mathbb{N}_n} \epsilon_i f(x_i) \\
&\leq C_\lambda^\psi \, \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{B}_\lambda} \sum_{i \in \mathbb{N}_n} \epsilon_i f(x_i) \right].
\end{aligned}$$

Also,

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{B}_\lambda} \sum_{i \in \mathbb{N}_n} \epsilon_i f(x_i) = \mathbb{E}_\varepsilon \quad \sqrt{\tfrac{1}{\lambda}} \sup_{K \in \mathcal{K}} \sup_{\|f\|_K \leq 1} \langle \sum_{i \in \mathbb{N}_n} \epsilon_i K_{x_i}, f \rangle_K)$$

$$\leq \sqrt{\tfrac{1}{\lambda}} \mathbb{E}_\varepsilon \sup_{K \in \mathcal{K}} \Big| \sum_{i,j \in \mathbb{N}_n} \varepsilon_i \varepsilon_j K(x_i, x_j) \Big|^{\frac{1}{2}}$$

$$\leq \sqrt{\tfrac{2n}{\lambda}} \sqrt{\hat{\mathcal{U}}_n(\mathcal{K})} + \sqrt{\tfrac{1}{\lambda}} \sup_{K \in \mathcal{K}} \sqrt{trace(\mathbf{K})}.$$

Putting all the above inequalities back into (19) yields that

$$\mathbb{E}\Big[\sup_{f \in \mathcal{B}_\lambda} \mathcal{E}^\psi(f) - \mathcal{E}_{\mathbf{z}}^\psi(f) \Big] \leq 2C_\lambda^\psi \sqrt{\frac{2\hat{\mathcal{U}}_n(\mathcal{K})}{\lambda n}} + 2C_\lambda^\psi \kappa \Big(\frac{1}{\lambda n}\Big)^{\frac{1}{2}} + 2M_\lambda^\psi \Big(\frac{\ln \frac{2}{\delta}}{2n}\Big)^{\frac{1}{2}},$$

where we used the fact that $\text{trace}(\mathbf{K}) \leq \kappa^2 n$. Putting this back into inequality (18) yields the desired result.

By similar arguments as above, we can prove the second statement. This completes the proof of the theorem. □

We are ready to prove Theorem 1.

Proof of Theorem 1: Recall that $f_{\mathbf{z}} \in \mathcal{B}_\lambda$ and note that ϕ is a normalized classifying loss. Then, applying Theorem 4 with $\psi = \phi$ implies inequality (5). □

4 Estimating the Rademacher Chaos Complexity

In this section we discuss how to estimate the Rademacher chaos complexity. First, parallel to the properties of Rademacher averages, it is useful to outline some properties of the Rademacher chaos complexity some of which may be interesting in its own right.

Proposition 1 *Let $F_1, \cdots, F_k$ and H be classes of real functions on $X \times X$. Then the following holds true.*

(a) If $F \subseteq H$ then $\hat{\mathcal{U}}_n(F) \leq \hat{\mathcal{U}}_n(H)$.

(b) $\hat{\mathcal{U}}_n(conv(F)) = \hat{\mathcal{U}}_n(F)$.

(c) For any $c \in \mathbb{R}$, $\hat{\mathcal{U}}_n(cF) = |c|\hat{\mathcal{U}}_n(F)$.

(d) $\hat{\mathcal{U}}_n(\sum_{i \in \mathbb{N}_k} F_i) \leq \sum_{i \in \mathbb{N}_k} \hat{\mathcal{U}}_n(F_i)$.

(e) For any $1 < q < p < \infty$, the Khinchin-type inequality holds true

$$\left(\mathbb{E}_\varepsilon \sup_{f \in F} |\hat{U}_f(\varepsilon)|^q \right)^{\frac{1}{q}} \leq \left(\mathbb{E}_\varepsilon \sup_{f \in F} |\hat{U}_f(\varepsilon)|^p \right)^{\frac{1}{p}} \leq \left(\frac{p-1}{q-1} \right) \left(\mathbb{E}_\varepsilon \sup_{f \in F} |\hat{U}_f(\varepsilon)|^q \right)^{\frac{1}{q}}$$

Proof Properties (a), (c), and (d) are directly from Definition 2 of the Rademacher chaos complexity. To prove Property (b), we note, for any $m \in \mathbb{N}$, $f_k \in F$, and $\{\lambda_k : k \in \mathbb{N}_m\}$ satisfying $\sum_k \lambda_k = 1$ and $\lambda_k \geq 0$, that

$$\left| \sum_{i,j,i<j} \varepsilon_i \varepsilon_j \sum_{k \in \mathbb{N}_m} \lambda_k f_k(x_i, x_j) \right| \leq \sum_{k \in \mathbb{N}_m} \lambda_k \left| \sum_{i<j} \varepsilon_i \varepsilon_j f_k(x_i, x_j) \right|$$
$$\leq \sup_{f \in F} \left| \sum_{i<j} \varepsilon_i \varepsilon_j f(x_i, x_j) \right|.$$

Since $\lambda_k, f_k \in F$ are arbitrary, $\hat{\mathcal{U}}_n(\mathrm{conv}(F)) \leq \hat{\mathcal{U}}_n(F)$. The reverse inequality is obvious which completes the proof of the desired Property (b). The last property is from Theorem 3.2.2 of De La Peña and Giné (1999). □

Now we are in a position to prove Theorem 2 using the standard *chaining arguments*. The estimation of the Rademacher chaos complexity by entropy integrals is a simple version of of maximal inequalities based on metric entropy (De La Peña and Giné, 1999, Chapter 5) which we give a proof for completeness. To this end, let $\mathcal{G}$ be a set of functions on $X \times X$ and $\mathbf{x} = \{x_i \in X : i \in \mathbb{N}_n\}$, define the l^2 empirical metric of two functions $f, g \in \mathcal{G}$ by

$$d_{\mathbf{x}}(f, g) = \left(\frac{1}{n^2} \sum_{i,j \in \mathbb{N}_n, i<j} |f(x_i, x_j) - g(x_i, x_j)|^2 \right)^{\frac{1}{2}}.$$

The empirical covering number $\mathcal{N}(\mathcal{G}, d_{\mathbf{x}}, \eta)$ is the smallest number of balls with radius η required to cover $\mathcal{G}$.

We begin with a useful lemma which deals with a finite class of homogeneous Rademacher chaos of order two.

Lemma 1 *Let $\{f_\ell : \ell \in \mathbb{N}_N\}$ be a finite class of functions on $X \times X$ and $\{\epsilon_i : i \in \mathbb{N}_n\}$ are independent Rademacher random variables. Consider the homogeneous Rademacher chaos process of order two $\{\hat{U}_{f_\ell}(\varepsilon) = \frac{1}{n} \sum_{i,j \in \mathbb{N}_n, i<j} \varepsilon_i \varepsilon_j f_\ell(x_i, x_j) : \ell \in \mathbb{N}_N\}$. Then, we have that*

$$\mathbb{E}\left[\max_{\ell \in \mathbb{N}_N} |\hat{U}_{f_\ell}(\varepsilon)| \right] \leq 2e \log(1 + N) \max_{\ell \in \mathbb{N}_N} \left(\frac{1}{n^2} \sum_{i<j} |f_\ell(x_i, x_j)|^2 \right)^{\frac{1}{2}},$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to the Rademacher variable ε.

14

Proof: By Jensen's inequality,

$$
\begin{aligned}
e^{\lambda \mathbb{E}\left[\max_{\ell \in \mathbb{N}_N} |\hat{U}_{f_\ell}(\varepsilon)|\right]} - 1 &\leq \mathbb{E}\left[e^{\lambda \max_{\ell \in \mathbb{N}_N} |\hat{U}_{f_\ell}(\varepsilon)|} - 1\right] \\
&= \mathbb{E}\left[\max_{\ell \in \mathbb{N}_N}\left(e^{\lambda |\hat{U}_{f_\ell}(\varepsilon)|} - 1\right)\right] \qquad (20) \\
&\leq \sum_{\ell \in \mathbb{N}_N} \mathbb{E}\left[\left(e^{\lambda |\hat{U}_{f_\ell}(\varepsilon)|} - 1\right)\right].
\end{aligned}
$$

For any $\ell \in \mathbb{N}_N$, the term $\mathbb{E}[e^{\lambda |\hat{U}_{f_\ell}(\varepsilon)|} - 1]$ can be estimated by the Khinchin-type inequality (see property (e) in Proposition 1) as follows:

$$
\begin{aligned}
\mathbb{E}[e^{\lambda |\hat{U}_{f_\ell}(\varepsilon)|} - 1] &= \sum_{k \geq 1} \frac{1}{k!} \lambda^k \mathbb{E}\left[|\hat{U}_{f_\ell}(\varepsilon)|^k\right] \\
&\leq \sum_{k \geq 1} \frac{1}{k!} \lambda^k k^k \left[\mathbb{E}|\hat{U}_{f_\ell}(\varepsilon)|^2\right]^{\frac{k}{2}} \qquad (21) \\
&\leq \sum_{k \geq 1} \left(e\lambda \left[\mathbb{E}|\hat{U}_{f_\ell}(\varepsilon)|^2\right]^{\frac{1}{2}}\right)^k.
\end{aligned}
$$

Here, in the second to last inequality of (21), we used the fact the $\mathbb{E}\left[|\hat{U}_{f_\ell}(\varepsilon)|\right] \leq \mathbb{E}\left[|\hat{U}_{f_\ell}(\varepsilon)|^2\right]^{\frac{1}{2}}$ and, for $k \geq 2$, the Khinchin-type inequality for homogeneous Rademacher chaos process of order two: $\mathbb{E}\left[|\hat{U}_{f_\ell}(\varepsilon)|^k\right] \leq k^k \left[\mathbb{E}|\hat{U}_{f_\ell}(\varepsilon)|^2\right]^{\frac{k}{2}}$. In the last inequality of (21), we used the Stirling's inequality i.e. $e^{-k} k^k \leq k!$.

Now set $\lambda = \left(2e \max_{\ell \in \mathbb{N}_N} \left[\mathbb{E}|\hat{U}_{f_\ell}(\varepsilon)|^2\right]^{\frac{1}{2}}\right)^{-1}$, the above inequality can be bounded by

$$
\mathbb{E}[e^{\lambda |\hat{U}_{f_\ell}(\varepsilon)|} - 1] \leq \sum_{k \geq 1} 2^{-k} = 1, \quad \forall \ell \in \mathbb{N}_N.
$$

Putting this back into (20) yields that

$$
e^{\lambda \mathbb{E}\left[\max_{\ell \in \mathbb{N}_N} |\hat{U}_{f_\ell}(\varepsilon)|\right]} - 1 \leq N.
$$

Equivalently,

$$
\mathbb{E}\left[\max_{\ell \in \mathbb{N}_N} |\hat{U}_{f_\ell}(\varepsilon)|\right] \leq 2e \log(1 + N) \max_{\ell \in \mathbb{N}_N} \left[\mathbb{E}|\hat{U}_{f_\ell}(\varepsilon)|^2\right]^{\frac{1}{2}}. \qquad (22)
$$

Observe that

$$
\mathbb{E}|\hat{U}_{f_\ell}(\varepsilon)|^2 = \frac{1}{n^2} \sum_{i<j, i'<j'} \mathbb{E}\left[\varepsilon_i \varepsilon_j \varepsilon_{i'} \varepsilon_{j'} f_\ell(x_i, x_j) f_\ell(x_{i'}, x_{j'})\right] = \sum_{i<j} f_\ell(x_i, x_j)^2 / n^2.
$$

Plugging this back into inequality (22) completes the proof of the lemma. $\qquad \square$

Equipped with the above lemma, we can prove Theorem 2 by the standard chaining arguments. To this end, let D be the diameter of $\mathcal{K}$ with respect to $d_\mathbf{x}$ then

$$
D = \sup_{K_1, K_2 \in \mathcal{K}} d_\mathbf{x}(K_1, K_2) \leq 2 \sup_{K \in \mathcal{K}} \left(\frac{1}{n^2} \sum_{i<j} |K(x_i, x_j)|^2\right)^{\frac{1}{2}} \leq 2\kappa^2.
$$

Proof of Theorem 2: For each $k = 0, 1, 2, \ldots$, let $\mathcal{K}^{(k)}$ be a minimal cover of $\mathcal{K}$ of radius $D2^{-k}$ and the cardinality of $\mathcal{K}^{(k)}$ denoted by $|\mathcal{K}^{(k)}| = \mathcal{N}(\mathcal{K}, d_{\mathbf{x}}, D2^{-k})$. Without loss of generality, choose some $K_0 \in \mathcal{K}$ and let $\mathcal{K}^{(0)} = \{K_0\}$. For any Rademacher variable ε, let

$$K^* = \arg \sup_{K \in \mathcal{K}} |\hat{U}_K(\varepsilon)|$$

and choose a $K_k^* \in \mathcal{K}^{(k)}$ whose distance to K^* is minimal. Obviously,

$$d_{\mathbf{x}}(K_{k-1}^*, K_k^*) \le d_{\mathbf{x}}(K_{k-1}^*, K^*) + d_{\mathbf{x}}(K^*, K_k^*) \le D2^{-(k-1)} + D2^{-k} = 3D2^{-k}. \quad (23)$$

Moreover, $\lim_{k \to \infty} d_{\mathbf{x}}(K^*, K_k^*) \to 0$. Hence,

$$\sup_{K \in \mathcal{K}} |\hat{U}_K(\varepsilon)| = |\hat{U}_{K^*}(\varepsilon)| = |\hat{U}_{K_0}(\varepsilon) + \sum_{k \in \mathbb{N}} (\hat{U}_{K_k^*}(\varepsilon) - \hat{U}_{K_{k-1}^*}(\varepsilon))|,$$

and therefore

$$
\begin{aligned}
\mathbb{E}\Big[\sup_{K \in \mathcal{K}} |\hat{U}_K(\varepsilon)|\Big] &\le \mathbb{E}\Big[|\hat{U}_{K_0}(\varepsilon)|\Big] + \sum_{k \in \mathbb{N}} \mathbb{E}\Big[\max_{\substack{(K, K') \in \mathcal{K}^{(k)} \times \mathcal{K}^{(k-1)} \\ d_{\mathbf{x}}(K, K') \le 3D2^{-k}}} |\hat{U}_K(\varepsilon) - \hat{U}_{K'}(\varepsilon)|\Big] \\
&\le \Big(\frac{1}{n^2} \sum_{i < j} |K_0(x_i, x_j)|^2\Big)^{\frac{1}{2}} + \sum_{k \in \mathbb{N}} \mathbb{E}\Big[\max_{\substack{(K, K') \in \mathcal{K}^{(k)} \times \mathcal{K}^{(k-1)} \\ d_{\mathbf{x}}(K, K') \le 3D2^{-k}}} |\hat{U}_{K-K'}(\varepsilon)|\Big].
\end{aligned}
$$

Applying Lemma 1, we have, for $k \ge 1$, that

$$
\begin{aligned}
\mathbb{E}\Big[\max_{\substack{(K, K') \in \mathcal{K}^{(k)} \times \mathcal{K}^{(k-1)} \\ d_{\mathbf{x}}(K, K') \le 3eD2^{-k}}} |\hat{U}_{K-K'}(\varepsilon)|\Big] &\le 3eD2^{-k} \log(1 + \mathcal{N}(\mathcal{K}, d_{\mathbf{x}}, D2^{-k})\mathcal{N}(\mathcal{K}, d_{\mathbf{x}}, D2^{-(k-1)})) \\
&\le 6eD2^{-k} \log(1 + \mathcal{N}(\mathcal{K}, d_{\mathbf{x}}, D2^{-k}))
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\mathbb{E}\Big[\sup_{K \in \mathcal{K}} |\hat{U}_K(\varepsilon)|\Big] &\le \kappa^2 + \sum_{k \ge 1} 6D2^{-k} \log(1 + \mathcal{N}(\mathcal{K}, d_{\mathbf{x}}, D2^{-k})) \\
&\le \kappa^2 + 12e \int_0^{D/2} \log(1 + \mathcal{N}(\mathcal{K}, d_{\mathbf{x}}, \delta)) d\delta.
\end{aligned}
$$

Combining this with the estimation $D \le 2\kappa^2$ completes the proof of Theorem 2. $\qquad \square$

It is worth mentioning that the above arguments hold true for the suprema of homogeneous Rademacher chaos processes of order m and a general function space F (not only the space of kernels). Here, the Rademacher chaos processes of order one is reduced to the standard Rademacher averages. The only difference in the proof is the Khinchin-type inequality. For instance, for the homogeneous Rademacher chaos processe $\{X_f : f \in F\}$ of order m, the general Khinchin-type inequality is given by

$$\Big(\mathbb{E}_\varepsilon \sup_{f \in F} |X_f(\varepsilon)|^q\Big)^{\frac{1}{q}} \le \Big(\mathbb{E}_\varepsilon \sup_{f \in F} |X_f(\varepsilon)|^p\Big)^{\frac{1}{p}} \le \Big(\frac{p-1}{q-1}\Big)^{\frac{m}{2}} \Big(\mathbb{E}_\varepsilon \sup_{f \in F} |X_f(\varepsilon)|^q\Big)^{\frac{1}{q}}.$$

16

By this inequality, we can show that, in analogy to the proof of Theorem 2, the superma of a homogeneous Rademacher chaos process of order m is bounded by the following entropy integral

$$\int_0^{D/2} \left[\log \mathcal{N}(\mathcal{K}, d_{\mathbf{x}}, \delta) \right]^{\frac{m}{2}} d\delta.$$

One can refer to Arcones and Giné (1993); De La Peña and Giné (1999) for more general entropy integrals to bound the suprema of Rademacher chaos processes of order m for any $m \in \mathbb{N}$.

Now we turn our attention to the proof of Theorem 3 in Section 2 which states that the empirical covering number is further estimated by the shattering dimension (Alon et al., 1997; Anthony and Bartlett, 1999) of the set of candidate kernels.

Proof of Theorem 3: For the first assertion, observe that the pseudo-dimension is equivalent to the VC-dimension of the following space (Anthony and Bartlett, 1999, Theorem 11.4)

$$\left\{ ((x, x'), t) \in X \times X \times \mathbb{R} : g((x, x'), t) = \text{sgn}(t - K(x, x')), \ K \in \mathcal{K} \right\}.$$

Combining this fact with (Bartlett, 2006, Theorem 3.1), we have

$$\mathcal{N}(\mathcal{K}, d_{\mathbf{x}}, \varepsilon) \le 2 \left(\frac{4e\kappa^4}{\varepsilon^2} \right)^{d_{\mathcal{K}}}, \tag{24}$$

which completes the proof of the first assertion[2].

For the second assertion, we obtain from Theorem 2 and inequality (24) that

$$
\begin{aligned}
\hat{\mathcal{U}}_n(\mathcal{K}) \ &\le \kappa^2 + 12e \int_0^{\kappa^2} \log\left(1 + \mathcal{N}(\mathcal{K}, d_{\mathbf{x}}, \varepsilon)\right) d\varepsilon \\
&\le \kappa^2 + 12e \int_0^{\kappa^2} \ln\left[e\left(\frac{4e\kappa^4}{\varepsilon^2}\right)^{d_{\mathcal{K}}} \right] d\varepsilon \\
&\le \kappa^2 + 12e\kappa^2 + 12e\ln(4e)\kappa^2 d_{\mathcal{K}} + 12e \int_0^{\kappa^2} \ln\left(\frac{\kappa^4}{\varepsilon^2}\right)^{d_{\mathcal{K}}} d\varepsilon
\end{aligned}
$$

Observe that

$$\int_0^{\kappa^2} \ln\left(\frac{\kappa^4}{\varepsilon^2}\right)^{d_{\mathcal{K}}} d\varepsilon = 2\kappa^2 d_{\mathcal{K}} \int_0^1 \ln\frac{1}{\varepsilon} d\varepsilon = 4\kappa^2 d_{\mathcal{K}}.$$

Putting these estimates together implies that

$$\hat{\mathcal{U}}(\mathcal{K}) \le (12e + 1)\kappa^2 + \kappa^2 (60e + 12e \ln 4)d_{\mathcal{K}} \le (96e + 1)\kappa^2 d_{\mathcal{K}},$$

---

[2]Similar covering number bound was also established in (van der Vaart and Wellner, 1996, Theorem 2.6.7): there exists a universal constant C such that $\mathcal{N}(\mathcal{K}, d_{\mathbf{x}}, \varepsilon) \le Cd_{\mathcal{K}}(16e)^{d_{\mathcal{K}}} \left(\frac{\kappa^2}{\varepsilon}\right)^{2(d_{\mathcal{K}}-1)}$. However, we failed to work out what is the universal constant C.

which completes the proof of the theorem. □

For the set of scalar Gaussian kernels given by equation (10), we have the following estimation.

Lemma 2 *Let the set of basis kernels $\mathcal{K}_{gau}$ be given by equation (10), then we have $d_{\mathcal{K}_{gau}} = 1$.*

Proof: It is obvious that there exists at least one pair of points $(x, t) \in X \times X$ such that it is pseudo-shattering by $\mathcal{K}$. Now assume that two pairs of points (x_1, t_1) and (x_2, t_2) are shattering by $\mathcal{K}$. By Definition 3 of pseudo-dimension, there exists $r_1, r_2 \in \mathbb{R}$ and $\sigma, \sigma' \in [0, \infty)$ such that

$$e^{-\sigma \|x_1 - t_1\|^2} > r_1, \quad e^{-\sigma \|x_2 - t_2\|^2} < r_2,$$

and

$$e^{-\sigma' \|x_1 - t_1\|^2} < r_1, \quad e^{-\sigma' \|x_2 - t_2\|^2} > r_2.$$

Hence,

$$e^{-\sigma \|x_1 - t_1\|^2} > e^{-\sigma' \|x_1 - t_1\|^2}, \text{ and } e^{-\sigma \|x_2 - t_2\|^2} < e^{-\sigma' \|x_2 - t_2\|^2}.$$

Equivalently, $\sigma < \sigma'$, and $\sigma > \sigma'$, which is obviously a contradiction. Consequently, the pseudo-dimension of $\mathcal{K}_{gau}$ is identical to one. □

We are ready to prove Corollary 1 with estimation of the Rademacher chaos complexities of $\mathcal{K}_{gau}$ and $\mathcal{K}_{rbf}$.

Proof of Corollary 1: The first statement follows directly from Theorem 2 and the observation that $\mathcal{N}(\mathcal{K}_{finite}, d_{\mathbf{x}}, \varepsilon) \leq m$ where m denotes the number of kernels in the set $\mathcal{K}_{finite}$.

Note that $\kappa = 1$. Then, the estimation of $\hat{\mathcal{U}}_n(\mathcal{K}_{gau})$ follows immediately by combining inequality (9) in Theorem 3 with Lemma 2. For the RBF kernels set $\mathcal{K}_{rbf}$, note, for any $\{x_i : i \in \mathbb{N}_n\}$, that

$$
\begin{aligned}
\hat{\mathcal{U}}_n(\mathcal{K}_{rbf}) &\leq \mathbb{E}_\varepsilon \sup_{p \in \mathcal{M}(\mathbb{R}^+)} \Big| \int_0^\infty \sum_{i<j} \varepsilon_i \varepsilon_j e^{-\sigma \|x_i - x_j\|^2} dp(\sigma) \Big| / n \\
&\leq \mathbb{E}_\varepsilon \sup_{\sigma \in \mathbb{R}^+} \Big| \sum_{i<j} \varepsilon_i \varepsilon_j e^{-\sigma \|x_i - x_j\|^2} \Big| / n \leq \hat{\mathcal{U}}_n(\mathcal{K}_{gau}).
\end{aligned}
$$

This completes the proof of the corollary. □

The estimation of pseudo-dimensions for Gaussian kernels with covariance matrices can be referred to Srebro and Ben-David (2006); Anthony and Bartlett (1999).

5   Deriving Error Rates in Classificaiton

We are now ready to derive explicit error rates for classification using the above generalization bounds. In subsequent examples we emphasize that the set of base kernels are given by either Gaussian kernels defined by equation (10) or the RBF kernels defined by equation (11).

We begin with the proofs of example 1 and example 2 stated in Section 2.1. To this end, we notice that, by the definition of f_λ^ϕ, we have $\mathcal{E}^\phi(f_\lambda^\phi) + \lambda \|f_\lambda^\phi\|_K^2 \leq \mathcal{E}(0) + \lambda \|0\|_K^2 = \mathcal{E}^\phi(0) = 1$ which implies that $\|f_\lambda^\phi\|_K \leq \sqrt{1/\lambda}$.

Proof of Example 1:   First note, for the hinge loss, that $C_\lambda^\phi = 1$ and $M_\lambda^\phi \leq 1 + \frac{\kappa}{\sqrt{\lambda}}$ and observe that $\mathcal{S}_{\mathbf{z},\lambda} \leq \sup_{f \in \mathcal{B}_\lambda} \left[\mathcal{E}^\phi(f) - \mathcal{E}_{\mathbf{z}}^\phi(f)\right] + \sup_{f \in \mathcal{B}_\lambda} \left[\mathcal{E}_{\mathbf{z}}^\phi(f) - \mathcal{E}^\phi(f)\right]$. Then, combining Theorem 4, Corollary 1 and the error decomposition (15) together, with probability at least $1 - \delta$ there holds that

$$\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}^\phi(f_c) \leq \mathcal{O}\left(\left(\frac{1}{n\lambda}\right)^{\frac{1}{2}} + \left(\frac{\ln \frac{4}{\delta}}{n\lambda}\right)^{\frac{1}{2}}\right) + \mathcal{D}(\lambda). \tag{25}$$

In addition, we know from Theorem 10 of Chen et al. (2004) that if the distribution enjoys the weakly separation condition with exponent θ then the regularization error decays as $\mathcal{D}(\lambda) = \mathcal{O}\left(\lambda^{\frac{\theta}{\theta+2}}\right)$. Letting $\lambda = n^{-\frac{\theta+2}{3\theta+2}}$. Combining inequality (25) with the comparison inequality (e.g. Bartlett et al. (2006); Zhang (2004))

$$\mathcal{R}(\text{sgn}(f_{\mathbf{z}}^\phi)) \leq \mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}^\phi(f_c)$$

yields the desired result.   □

Proof of Corollary 2: The margin-based cost function ψ obviously satisfies the conditions in Theorem 4 with $C_\lambda^\psi = \frac{1}{\gamma}$ and $M_\lambda^\psi = 1$. Since $\chi_{y \neq \text{sgn}(f(x))} \leq \psi(yf(x))$, there holds that $\mathcal{R}(\text{sgn}(f_{\mathbf{z}}^\phi)) \leq \mathcal{E}^\psi(f_{\mathbf{z}}^\phi)$ which, combining with inequality (5) in Theorem 1, yields the desired assertion.   □

Proof of Example 2: The results can be directly obtained by combining Corollary 1 with Corollary 2.   □

Now we turn our attention to general q-norm soft margin SVM losses $\phi(t) = (1 - t)_+^q$ for $q \in (1, \infty)$ for classification. In this case, we know from Chen et al. (2004) that the target function f_ρ^ϕ becomes

$$f_\rho^\phi(x) = f_q(x) = \frac{(1 + f_\rho(x))^{\frac{1}{q-1}} - (1 - f_\rho(x))^{\frac{1}{q-1}}}{(1 + f_\rho(x))^{\frac{1}{q-1}} + (1 - f_\rho(x))^{\frac{1}{q-1}}},$$

where $f_\rho(x) := P(Y = 1|x) - P(Y = -1|x)$.

Example 3 *Let $\phi(t) = (1 - t)_+^q$ for some $q \in (1, \infty)$ and suppose that the separation condition holds true with exponent $\theta > 0$. Then, choosing $\lambda = n^{-\frac{q\theta}{4+2(2q+1)\theta}}$ with probability at least $1 - \delta$ there holds*

$$\mathcal{R}(sgn(f_{\mathbf{z}}^\phi)) - \mathcal{R}(f_c) \leq \mathcal{O}\left(\left[\ln \frac{1}{\delta}\right]^{\frac{1}{4}} n^{-\frac{q\theta}{4+2(2q+1)\theta}}\right)$$

Proof: First observe that $C_\lambda^\phi \leq (1 + \frac{1}{\sqrt{\lambda}})^{q-1}$ and $M_\lambda^\phi \leq (1 + \frac{\kappa}{\sqrt{\lambda}})^q$. Hence, from Theorem 4, Corollary 1 and the error decomposition (15), we know, for any $\lambda \in (0, 1)$, that

$$\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}^\phi(f_q) \leq \mathcal{O}\left(\left(\frac{1}{n\lambda^q}\right)^{\frac{1}{2}} + \left(\frac{\ln \frac{4}{\delta}}{n\lambda^q}\right)^{\frac{1}{2}}\right) + \mathcal{D}(\lambda).$$

Also, we know from (Chen et al., 2004, Theorem 10) that if the distribution enjoys the weakly separation condition with exponent θ then the regularization error decays as $\mathcal{D}(\lambda) = \mathcal{O}\left(\lambda^{\frac{\theta}{\theta+2}}\right)$. Letting $\lambda = n^{-\frac{q(\theta+2)}{2+(2q+1)\theta}}$ yields that

$$\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}^\phi(f_q) \leq \mathcal{O}\left(\left[\ln \frac{1}{\delta}\right]^{\frac{1}{2}} n^{-\frac{q\theta}{2+(2q+1)\theta}}\right).$$

Recall the comparison inequality (Theorem 14 of Chen et al. (2004)) for q-norm SVM:

$$\mathcal{R}(\text{sgn}(f_{\mathbf{z}}^\phi)) - \mathcal{R}(f_c) \leq \sqrt{2\left(\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}^\phi(f_q)\right)}.$$

Consequently, with probability at least $1 - \delta$ there holds

$$\mathcal{R}(\text{sgn}(f_{\mathbf{z}}^\phi)) - \mathcal{R}(f_c) \leq \mathcal{O}\left(\left[\ln \frac{1}{\delta}\right]^{\frac{1}{4}} n^{-\frac{q\theta}{4+2(2q+1)\theta}}\right),$$

which completes the proof of the example. □

Our last example is the least square loss for classification which is extensively studied in the single kernel case (Caponnetto and De Vito, 2007; De Vito et al., 2006; Smale

and Zhou, 2004; Zhang, 2004). In this case, in order to get meaningful rates of the regularization error $\mathcal{D}(\lambda)$ we can assume the target function enjoys some Sobolev smoothness. Recall in the regression case, the target function $f_\rho^\phi = f_\rho(x)$ for any $x \in X$ usually referred to as the *regression function* and the nature of least square loss implies that

$$\mathcal{E}(f_\mathbf{z}^\phi) - \mathcal{E}(f_\rho) = \int_X |f_\mathbf{z}^\phi(x) - f_\rho(x)|^2 d\rho_X(x).$$

Example 4 *Let X be a domain in $\mathbb{R}^d$ with Lipschitz boundary. Assume the regression function $f_\rho \in H^s(X)$ with some $s > 0$. Then the following holds true.*

1. *If $d/2 < s \leq d/2 + 2$ then for any $0 < \varepsilon < 2s - d$, by taking $\lambda = n^{-\frac{2s-\varepsilon}{2(4s-d-2\varepsilon)}}$, with probability at least $1 - \delta$ there holds*

$$\begin{aligned}\mathcal{R}(sgn(f_\mathbf{z}^\phi)) - \mathcal{R}(f_c) &\leq \left(\int_X |f_\mathbf{z}^\phi(x) - f_\rho(x)|^2 d\rho_X(x) \right)^{\frac{1}{2}} \\ &\leq \mathcal{O}\left(\left[\ln \tfrac{1}{\delta} \right]^{\frac{1}{4}} n^{-\frac{2s-d-\varepsilon}{4(4s-d-2\varepsilon)}} \right).\end{aligned}$$

2. *If X is bounded, ρ_X is the Lebesgue measure, and $0 < s \leq 2$ then by choosing $\lambda = n^{-\frac{2s+d}{2(4s+d)}}$, with probability at least $1 - \delta$, there holds*

$$\begin{aligned}\mathcal{R}(sgn(f_\mathbf{z}^\phi)) - \mathcal{R}(f_c) &\leq \left(\int_X |f_\mathbf{z}^\phi - f_\rho|^2 d\rho_X(x) \right)^{\frac{1}{2}} \\ &\leq \mathcal{O}\left(\left[\ln \tfrac{1}{\delta} \right]^{\frac{1}{4}} n^{-\frac{s}{2(4s+d)}} \right).\end{aligned}$$

Proof: For the least square loss, we observe that $C_\lambda^\phi = 2(1 + \frac{1}{\sqrt{\lambda}})$ and $M_\lambda^\phi \leq (1 + \frac{\kappa}{\sqrt{\lambda}})^2$. Then, we know from Theorem 4, Corollary 1 and the error decomposition (15) that

$$\begin{aligned}\mathcal{E}(f_\mathbf{z}^\phi) - \mathcal{E}(f_\rho) &= \int_X |f_\mathbf{z}^\phi(x) - f_\rho(x)|^2 d\rho_X(x) \\ &\leq \mathcal{O}\left(\left(\tfrac{1}{n\lambda^2} \right)^{\frac{1}{2}} + \left(\tfrac{\ln \frac{2}{\delta}}{n\lambda^2} \right)^{\frac{1}{2}} + \tfrac{1}{\sqrt{n}} \right) + \mathcal{D}(\lambda).\end{aligned} \tag{26}$$

Then, for the first assertion we know from Proposition 22 of Ying and Zhou (2007) that

$$\mathcal{D}(\lambda) \leq \mathcal{O}\left(\lambda^{\frac{2s-\varepsilon-d}{2s-\varepsilon}} \right).$$

Putting the above two equations together and letting $\lambda = n^{-\frac{2s-\varepsilon}{2(4s-2\varepsilon-d)}}$ implies that

$$\int_X |f_\mathbf{z}^\phi(x) - f_\rho(x)|^2 d\rho_X(x) \leq \mathcal{O}\left(\left[\ln \tfrac{1}{\delta} \right]^{\frac{1}{2}} n^{-\frac{2s-d-\varepsilon}{2(4s-d-2\varepsilon)}} \right).$$

Hence, the desired result follows from the comparison inequality (Chen et al., 2004; Bartlett et al., 2006; Zhang, 2004) for the least square loss:

$$\mathcal{R}(\text{sign}(f_{\mathbf{z}}^{\phi})) - \mathcal{R}(f_c) \leq \sqrt{2\Big(\mathcal{E}^{\phi}(f_{\mathbf{z}}^{\phi}) - \mathcal{E}^{\phi}(f_{\rho})\Big)}. \tag{27}$$

The proof of the second assertion is similar as above. Recall that Proposition 22 of Ying and Zhou (2007) implies that the regularization error is estimated as follows:

$$\mathcal{D}(\lambda) \leq \mathcal{O}\Big(\lambda^{\frac{2s}{2s+d}}\Big).$$

Combining this with inequality (26) and the comparison inequality (27), with choice $\lambda = n^{-\frac{2s+d}{2(4s+d)}}$ we get the desired second assertion. □

We end this section with a comparison with error rates in Ying and Zhou (2007) on the least square loss for classification. In Example 1 there, it was proven that: if $d/2 < s \leq d/2 + 2$ then for any $0 < \varepsilon < 2s - d$, we have that

$$\mathbb{E}\Big[\int_X |f_{\mathbf{z}}^{\phi}(x) - f_{\rho}(x)|^2 d\rho_X(x)\Big]^{\frac{1}{2}} \leq \Big(\mathbb{E}\Big[\int_X |f_{\mathbf{z}}^{\phi}(x) - f_{\rho}(x)|^2 d\rho_X(x)\Big]\Big)^{\frac{1}{2}}$$
$$\leq \mathcal{O}\Big(n^{-\frac{2s-d-\varepsilon}{8(4s-d-2\varepsilon)}}\Big).$$

Ignoring the difference of the forms to express error rates using expectations and probabilistic inequalities, Example 4 yields that $\mathcal{O}\Big(n^{-\frac{2s-d-\varepsilon}{4(4s-d-2\varepsilon)}}\Big)$. Likewise, for the case $0 < s \leq 2$ and ρ_X is the Lebesgue measure, we got improved rates $\mathcal{O}\Big(n^{-\frac{s}{2(4s+d)}}\Big)$ in comparison with $\mathcal{O}\Big((\ln n)^{\frac{1}{4}} n^{-\frac{s}{4(4s+d)}}\Big)$ obtained previously. Hence, our new error rates substantially improve those in Ying and Zhou (2007).

6 Related Work and Discussion

Statistical bounds with Rademacher complexities were first pursued by Lanckriet et al. (2004); Bousquet and Herrmann (2003) for learning the kernel from a linear combination of finite candidate kernels. The Rademacher complexities are estimated by the eigenvalues of the candidate kernel matrix over the inputs.

Ying and Zhou (2007) pioneered the generalization analysis of learning Gaussians with varying variances. In particular, it was proved the union space $\mathcal{B}_{\mathcal{K}}$ is a uniform Glivenko-Cantelli (uGC) class (see definition in Alon et al. (1997)) if and only if, for any $\gamma > 0$, the V_{γ}-dimension of $\mathcal{K}_X = \{K(\cdot, x) : x \in X, K \in \mathcal{K}\}$ is finite. There,

22

the empirical covering number of $\mathcal{K}_X$ for Gaussians was also estimated. Based on these main results, the Rademacher bounds were established in Ying and Zhou (2007); Micchelli et al. (2005)[3]:

$$\mathcal{E}(f_{\mathbf{z}}^\phi) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^\phi) \leq 4C_\lambda^\phi \left(\frac{2R_n(\mathcal{K}_X)}{\sqrt{n}\lambda}\right)^{\frac{1}{2}} + 4\kappa C_\lambda^\phi \left(\frac{1}{\sqrt{n}\lambda}\right)^{\frac{1}{2}}$$
$$+ M_\lambda^\phi \left(\frac{\ln(\frac{2}{\delta})}{n}\right)^{\frac{1}{2}} + \frac{2}{\sqrt{n}}.$$

Here, the Rademacher complexity $R_n(\mathcal{K}_X)$ is defined by $\mathbb{E}\sup_{f\in\mathcal{K}_X}\frac{1}{\sqrt{n}}\left|\sum_{i\in\mathbb{N}_n}\varepsilon_i f(x_i)\right|$ which is often bounded by $\mathcal{O}(d_\mathcal{K}\ln n)$ by using metric entropy integrals, see Theorem 20 in Ying and Zhou (2007). Hence, the resultant rates are quite loose whose dependence on the sample number is of order $n^{-\frac{1}{4}}$ in comparison with our new bound of order $n^{-\frac{1}{2}}$ summarized in equation (13). Specifically, for the hinge loss, as stated in Example 1 we can get a better rate $\mathcal{O}(n^{-\frac{\theta}{2+3\theta}})$ in comparison with the rate $\mathcal{O}((\log n)^{\frac{1}{2}}n^{-\frac{\theta}{2(2+3\theta)}})$ given in Ying and Zhou (2007).

Srebro and Ben-David (2006) employed matrix analysis techniques to directly estimate the empirical covering number of $\mathcal{B}_\mathcal{K}$ with the pseudo-dimension of the candidate kernels. Margin bounds were established for SVM. Specifically, let $\mathcal{R}_{\mathbf{z}}^\gamma(f) = \frac{|\{i:y_i f(x_i)<\gamma\}|}{n}$. Note $f_{\mathbf{z}}^\phi \in \frac{1}{\sqrt{\lambda}}\mathcal{B}_\mathcal{K}$ where $\mathcal{B}_\mathcal{K}$ is the same as the notation $\mathcal{F}_\mathcal{K}$ used in Srebro and Ben-David (2006). A simple modification of Theorem 2 in Srebro and Ben-David (2006) to the function class $\frac{1}{\sqrt{\lambda}}\mathcal{B}_\mathcal{K}$, for any margin cost function ψ defined by equation (17), there holds

$$\mathcal{R}(\text{sgn}(f_{\mathbf{z}}^\phi)) \leq \mathcal{R}_{\mathbf{z}}^\gamma(f_{\mathbf{z}}^\phi) + \left(8(2+d_\mathcal{K})\ln\frac{128en^3\kappa^2}{\gamma^2\lambda d_\mathcal{K}} + 256\frac{\kappa^2}{\gamma^2\lambda}\ln\frac{128n\kappa^2}{\gamma^2\lambda} + \ln\frac{1}{\delta}\right)^{\frac{1}{2}}/\sqrt{n}.$$

Since $\mathcal{R}_{\mathbf{z}}^\gamma(f_{\mathbf{z}}^\phi) \geq \mathcal{E}_{\mathbf{z}}^\psi(f_{\mathbf{z}}^\phi)$, Corollary 2 implies

$$\mathcal{R}(\text{sgn}(f_{\mathbf{z}}^\phi)) \leq \mathcal{R}_{\mathbf{z}}^\gamma(f_{\mathbf{z}}^\phi) + 2\left(\frac{(192e+2)\kappa^2 d_\mathcal{K}}{n\lambda\gamma^2}\right)^{\frac{1}{2}} + 2\kappa\left(\frac{1}{n\lambda\gamma^2}\right)^{\frac{1}{2}} + 3\left(\frac{\ln\frac{2}{\delta}}{n}\right)^{\frac{1}{2}}.$$

Comparing the above two margin bounds, there is no logarithmic margin term, i.e. $\ln\frac{1}{\gamma^2}$, in our bound. The empirical covering approach Srebro and Ben-David (2006) is is roughly of the form $\left(d_\mathcal{K}\ln\frac{n}{\gamma^2} + \frac{1}{\gamma^2}\ln\frac{n}{\gamma^2}\right)^{\frac{1}{2}}/\sqrt{n}$. The Rademacher approach is of the form $\sqrt{\frac{d_\mathcal{K}}{n\gamma^2}}$ due to the contraction inequality of Rademacher averages for the margin

---

[3]This bound is originally given in the form of expectation. However, it is easy to convert it to the current probabilistic form by the bounded difference inequality from which the extra term $M_\lambda^\phi\left(\ln(\frac{1}{\delta})/n\right)^{\frac{1}{2}}$ appears.

cost function. Hence, our bound is comparable to their bounds. Moreover, there is no logarithmic term, i.e. $\ln n$, in our bound.

We can use the covering number in Srebro and Ben-David (2006) to derive generalization bounds. To see this, using standard symmetrization techniques and McDiarmid's inequality we have, with probablity $1 - \delta$, that

$$
\begin{aligned}
\mathcal{E}(f_{\mathbf{z}}^{\phi}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^{\phi}) \ &\leq 2\frac{R_n(\phi \circ \mathcal{B}_\lambda)}{\sqrt{n}} + M_\lambda^\phi \left(\frac{\ln(\frac{2}{\delta})}{n}\right)^{\frac{1}{2}} \\
&\leq 2C_\lambda^\phi \frac{R_n(\mathcal{B}_{\mathcal{K}})}{\sqrt{n\lambda}} + M_\lambda^\phi \left(\frac{\ln(\frac{2}{\delta})}{n}\right)^{\frac{1}{2}},
\end{aligned}
$$

where $\phi \circ \mathcal{B}_\lambda = \{\phi(yf(x)) : f \in \mathcal{B}_\lambda\}$. To estimate the Rademacher complexity, recall the scaling version of Theorem 1 in Srebro and Ben-David (2006):

$$
\mathcal{N}_n(\mathcal{F}_{\mathcal{K}}, \varepsilon) \leq 2\left(\frac{4en^3\kappa^2}{\varepsilon d_{\mathcal{K}}}\right)^{d_{\mathcal{K}}} \left(\frac{16n\kappa^2}{\varepsilon^2\lambda}\right)^{\frac{64\kappa^2}{\varepsilon^2}\ln\left(\frac{\varepsilon e n}{8\kappa}\right)}.
$$

Then, we use the following Dudley's entropy bound, for any $N \in \mathbb{N}$, there exists an absolute constant C such that for every $N \in \mathbb{N}$,

$$
R_n(\mathcal{B}_{\mathcal{K}}) \leq C\sum_{k=1}^{N} \varepsilon_{k-1}\log^{\frac{1}{2}} \mathcal{N}(\mathcal{F}_{\mathcal{K}}, d_{\mathbf{X}}, \varepsilon_k) + 2\varepsilon_N n^{\frac{1}{2}}.
$$

Since $\mathcal{N}(\mathcal{F}_{\mathcal{K}}, d_{\mathbf{X}}, \varepsilon_k) \leq \mathcal{N}_n(\mathcal{F}_{\mathcal{K}}, \varepsilon_k)$, selecting $\varepsilon_k = 2^{-k}$ and $N = \frac{\log n}{2}$ implies that $R_n(\mathcal{B}_{\mathcal{K}}) \leq Cd_{\mathcal{K}}(\ln n)^{\frac{3}{2}}$. Hence,

$$
\mathcal{E}(f_{\mathbf{z}}^{\phi}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^{\phi}) \leq C\frac{d_{\mathcal{K}}^{\frac{1}{2}}(\ln n)^{\frac{3}{2}}}{\sqrt{n\lambda}} + M_\lambda^\phi \left(\frac{\ln(\frac{2}{\delta})}{n}\right)^{\frac{1}{2}} + \frac{2}{\sqrt{n}}.
$$

In contrast, our generalization bound given by (13) is slightly better since it mainly depends on $\sqrt{\frac{d_{\mathcal{K}}}{n\lambda}}$. Moreover, Rademacher approaches are usually more flexible. For instance, it is unknown how to directly estimate the pseudo-dimension of RBF kernels $\mathcal{K}_{\mathrm{rbf}}$ and hence it could be a problem to directly apply the approach of Srebro and Ben-David (2006). The Rademacher approaches can handle this general case using the Rademacher chaos complexity of $\mathcal{K}_{\mathrm{gau}}$ instead of directly using that of $\mathcal{K}_{\mathrm{rbf}}$ as stated in Corollary 1 in Section 2.

7 Conclusion

In this paper we provided a novel statistical generalization bound for kernel learning system which extends and improves previous work in the literature (Lanckriet et al.,

2004; Wu et al., 2006; Ying and Zhou, 2007; Micchelli et al., 2005; Srebro and Ben-David, 2006). The main tools are based on the theory of U-processes such as the so-called homogeneous Rademacher chaos of order two and metric entropy integrals involving empirical covering numbers. There are several questions remaining to be further studied.

- Firstly, it would be interesting to get fast error rates with respect to the sample number as those in Bartlett et al. (2006); Steinwart and Scovel (2005); Wu et al. (2006). For this purpose, the extension of localized Rademacher averages (Bartlett et al., 2005) to the scenario of multiple kernel learning would be useful.

- Secondly, it would be interesting to investigate generalization bounds based on decoupling Gaussian chaos of order two, see its definition in De La Peña and Giné (1999).

- Thirdly, as mentioned in Section 6, it remains unknown how to get additive margin bounds using Rademacher approaches.

- Finally, another direction for further investigation is to apply Rademacher Chaos complexities to practical kernel learning problems.

Acknowledgemts

References

N. Alon, S. Ben-David, S. N. Cesa-Bianchi, and D. Haussler. (1997). Scale-sensitive dimensions, uniform convergence and learnability. *Journal of the ACM*, **44**: 615–631.

Anthony M. and Bartlett P. L. (1999). *Neural Networks Learning: Theoretical Foundations*, Cambridge University Press.

Arcones M. A. and Giné E. (1993). Limit theorems for U-processes, *The Annals of Probability*, *21*: 1494–1542.

Bach F. (2008). Consistency of the group Lasso and multiple kernel learning, *J. of Machine Learning Research*, *9*, 1179-1225.

Bartlett P.L. (2006). Lecture notes on the course "Statistical Learning Theory", Lecture 25, Spring term, 2006. http://www.cs.berkeley.edu/~bartlett/courses/281b-sp06/

Bartlett P. L., Bousquet O., and Mendelson S. (2005). Local Rademacher complexities. *The Annals of Statistics*, **33**: 1497–1537.

Bartlett P. L., Jordan M. I., and McAuliffe J. D. (2006). Convexity, classification, and risk bounds, *J. of the American Statistical Association*, *473*: 138–156.

Bartlett P. L. and Mendelson S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *J. of Machine Learning Research*, *3*, 463–482.

Bousquet O. and Herrmann D.J.L. (2003). On the complexity of learning the kernel matrix. *Advances in Neural Information Processing Systems (NIPS)*.

Caponnetto A. and De Vito E. (2007). Optimal Rates for Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, *7*, 331–368.

Chen D.R., Wu Q., Ying Y., and Zhou D. X. (2004). Support vector machine soft margin classifiers: error analysis. *J. of Machine Learning Research*, *5*, 1143–1175.

Chapelle O., Vapnik V., Bousquet O., and Mukherjee S. (2002). Choosing multiple parameters for support vector machines, *Machine Learning*, *6*, 131–159.

Clémencon S., Lugosi G., and Vayatis N. (2008). Ranking and empirical minimization of U-statistics, *Annals of Statistics*, *36*, 844–874.

Cucker F. and Zhou D.X. (2007). *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge.

Damoulas T and Girolami M. (2008). Probabilistic multi-class multi-kernel learning: On protein fold recognition and remote homology detection, *Bioinformatics*, *24* (10), 1264-1270.

De La Peña V. H. and Giné E. (1999). *Decoupling: from Dependence to Independence*. Springer, New York.

De Vito E., Caponnetto A., and Rosasco L. (2006). Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, *5*, 59–85.

Devroye L., Györfi L., and Lugosi G. (1997). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

Girolami M. and Rogers S. (2005). Hierarchic Bayesian models for kernel learning, *Proceedings of the International Conference on Machine Learning (ICML)*.

Koltchinskii V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, *47*, 1902–1914.

Koltchinskii V. and Panchenko V. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, *30*, 1–50.

Lanckriet G. R. G., Cristianini N., Bartlett P. L., Ghaoui L. E., and Jordan M. I. (2004). Learning the kernel matrix with semidefinite programming. *J. of Machine Learning Research*, *5*, 27–72.

Mendelson S. (2002). Rademacher averages and phase transitions in Glivenko-Cantelli classes, *IEEE Transactions on Information Theory*, 48(1), 251-263.

Mendelson S. (2003). A few notes on Statistical Learning Theory. *Advanced Lectures in Machine Learning*, (S. Mendelson, A.J. Smola Eds), LNCS 2600: 1-40.

Micchelli C. A. and Pontil M. (2005). Learning the kernel function via regularization, *J. of Machine Learning Research*, *6*, 1099–1125.

Micchelli C. A., Pontil M., Wu Q., and Zhou D. X. (2005). Error bounds for learning the kernel, Technical Report, City University of Hong Kong.

Ong C. S., Smola A.J. and Williamson R.C. (2005). Learning the kernel with hyperkernels. *J. of Machine Learning Research 6*, 1043–1071.

Schoenberg I.J. (1938). Metric spaces and completely monotone functions, *Ann. of Math. 39*, 811–841.

Schölkopf B. and Smola A. J. (2002). *Learning with Kernels*. The MIT Press, Cambridge, MA, USA.

Smale S. and Zhou D. X. (2004). Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, *41*, 279–305.

Srebro N. and Ben-David S. (2006). Learning bounds for support vector machines with learned kernels. *Proceedings of 19th Annual Conference on Learning Theory (COLT)*.

Shawe-Taylor J. and Cristianini N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge university press.

Sonnenburg S., Rätsch G., Schäfer C., and Schölkopf B. (2006). Large scale multiple kernel learning. *J. of Machine Learning Research*, *7*, 1531–1565.

Steinwart I. and Scovel C. (2005). Fast rates for support vector machines. *Proceedings of 18th Annual Conference on Learning Theory (COLT)*.

Steinwart I. and Scovel C. (2007). Fast rates for support vector machines using gaussian kernels. *Annals of Statistics*, *35*, 575-607.

Steinwart I. and Christmann A. (2008). *Support Vector Machines*. Springer, New york.

Wu Q., Ying Y., and Zhou D. X. (2006). Multi-kernel regularized classifiers. *Journal of Complexity*, 2006.

Van der Vaart A. W. and Wellner J. A. (1996). *Weak Convergence and Empirical Processes*, Springer-Verlag.

Ye J., Ji S., and Chen J. (2008). Multi-class discriminant kernel learning via convex programming, *J. of Machine Learning Research*, *9*, 719–758.

Ye G. B. and Zhou D. X. (2008). Learning and approximation by Gaussians on Riemannian manifolds, *Adv. Comput. Math.*, *29* (3): 291–310.

Ying Y., Huang K. and Campbell C. (2009). Enhanced protein fold recognition through a novel data integration approach. *BMC Bioinformatics*, *10*: 267.

Ying Y. and Zhou D. X. (2007). Learnability of Gaussians with flexible variances, *J. of Machine Learning Research*, *8*, 249-276.

Ying Y. and Campbell C. (2009). Generalization bounds for learning the kernel. *Proceedings of 22nd Annual Conference on Learning Theory (COLT)*.

Zhang T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, *32*, 56-85.

Meir R. and Zhang T. (2003). Generalization error bounds for Bayesian mixture algorithms. *J. of Machine Learning Research*, **4**: 839–860, 2003.